

Biost 518 Applied Biostatistics II

Midterm Examination Key February 2, 2007

Name: _____ Disc Sect: M W F

Instructions: Please provide concise answers to all questions. Rambling answers touching on topics not directly relevant to the question will tend to count against you. Nearly telegraphic writing style is permissible.

The examination is closed book and closed notes. If you come to a problem that you believe cannot be answered without making additional assumptions, clearly state the reasonable assumptions that you make, and proceed.

- (For all calculations in this problem, please use at least 4 significant digits.) Suppose we are interested in the association between serum cholesterol, age, and body mass index ($bmi = \text{weight} / \text{height}^2$). The following are the results of a linear regression analysis of data on 5,000 elderly Americans. The variable definitions are

- *cholest*: serum cholesterol in mg/dl
- *age*: age in years
- *bmi*: body mass index in kg / m^2

```
. tabstat cholest age bmi, col(stat) stat(n mean sd min p25 p50 p75 max) format
```

variable	N	mean	sd	min	p25	p50	p75	max
cholest	4953	211.7	39.3	73.0	186.0	210.0	236.0	430.0
age	5000	72.8	5.6	65.0	68.0	72.0	76.0	100.0
bmi	4987	26.7	4.7	14.7	23.5	26.1	29.2	58.8

```
. regress cholest bmi age, robust
```

Linear regression

```
Number of obs = 4940
F( 2, 4937) = 16.39
Prob > F = 0.0000
R-squared = 0.0067
Root MSE = 39.17
```

cholest	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
bmi	.1589302	.1194685	1.33	0.183	-.0752812	.3931416
age	-.5431455	.1017397	-5.34	0.000	-.7426005	-.3436906
_cons	246.9779	8.439044	29.27	0.000	230.4337	263.5222

- a. (5 points) Based on the above regression model, what is the best estimate for the mean cholesterol in 70 year old subjects with a BMI of 25 kg/m²?

Ans: $247.0 + 70 \times (-0.5431) + 25 \times 0.1589 = 213.0 \text{ mg/dl}$

- b. (5 points) Based on the above regression model, what is the best estimate for the mean cholesterol in 71 year old subjects with a BMI of 25 kg/m²?

Ans: $247.0 + 71 \times (-0.5431) + 25 \times 0.1589 = 212.4 \text{ mg/dl}$ (= 213.0 - 0.5431 based on the answer to part a and the slope for age)

- c. (5 points) Based on the above regression model, what is the best estimate for the mean cholesterol in 70 year old subjects with a BMI of 26 kg/m²?

Ans: $247.0 + 70 \times (-0.5431) + 26 \times 0.1589 = 213.1 \text{ mg/dl}$ (= 213.0 + 0.1589 based on the answer to part a and the slope for bmi)

- d. (5 points) Based on the above regression model, what is the best estimate for the mean cholesterol in 80 year old subjects with a BMI of 25 kg/m²?

Ans: $247.0 + 80 \times (-0.5431) + 25 \times 0.1589 = 207.5 \text{ mg/dl}$ (= 213.0 - 10 × 0.5431 based on the answer to part a and the slope for age)

- e. (5 points) Based on the above regression model, what is the best estimate for the difference in mean cholesterol between 70 year old subjects with a BMI of 26 kg/m² and 70 year old subjects with a BMI of 25 kg/m²?

Ans: 0.1589 mg/dl (with the larger person tending toward the higher average cholesterol level) (This is just the slope for bmi in the regression model.)

- f. (5 points) Based on the above regression model, what is the best estimate for the difference in mean cholesterol between 80 year old subjects with a BMI of 25 kg/m² and 70 year old subjects with a BMI of 25 kg/m²?

Ans: $10 \times (-0.5431) = -5.431 \text{ mg/dl}$ (the older person tends toward the lower cholesterol level) (=207.5 - 213.0 based on the answers to part a and part d and roundoff error)

- g. (5 points) Based on the above regression model, what is the best estimate for the difference in mean cholesterol between two groups of subjects having the same BMI, but who differ in age by 10 years? Provide a confidence interval for this estimate.

Ans: $10 \times (-0.5431) = -5.431 \text{ mg/dl}$ (the older person tends toward the lower cholesterol level), with a 95% CI also found by multiplying the CI for the age slope parameter: -7.426 mg/dl to -3.437 mg/dl

- h. (5 points) Provide an interpretation for the intercept in the above regression model. What scientific use would you make of this estimate?

Ans: The intercept is the estimated mean cholesterol in newborns having no weight. There are no such people, and thus this number is not scientifically relevant. (One student suggested this meant invisible newborns, but as we cannot define 0 divided by 0, we must have newborns that are a line—height but no transverse area. If they existed, maybe we could see them.)

- i. (5 points) Provide an interpretation for the slope for the age predictor in the above regression model. What scientific use would you make of this estimate?

Ans: The slope parameter is the estimated difference in mean cholesterol between two populations differing in their ages by 1 year but agreeing in their BMI. This estimates the association between age and cholesterol after adjusting for BMI.

- j. (5 points) Is there evidence that the slope for the age predictor is different from 0? State your evidence.

Ans: Yes, the P value testing for a nonzero age slope parameter is $P < 0.005$, which would be judged statistically significant at the 0.05 level. (Equivalently, the 95% CI also does not contain 0.)

- k. (5 points) Provide an interpretation for the slope for the bmi predictor in the above regression model. What scientific use would you make of this estimate?

Ans: The slope parameter is the estimated difference in mean cholesterol between two populations differing in their BMI by 1 kg/m² but agreeing in their age. This estimates the association between BMI and cholesterol after adjusting for age.

- l. (5 points) Is there evidence that the slope for the bmi predictor is different from 0? State your evidence.

Ans: No, the P value testing for a nonzero BMI slope parameter is $P = 0.183$, which would not be judged statistically significant at the 0.05 level. (Equivalently, the 95% CI does contain 0.)

- 2. Now suppose we consider a log transformation of cholesterol: $\logchol = \log(\text{cholest})$. Consider the following linear regression analysis.

. regress logchol age bmi, robust

Linear regression

Number of obs = 4940
 F(2, 4937) = 17.79
 Prob > F = 0.0000
 R-squared = 0.0077
 Root MSE = .18808

logchol	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
age	-.002828	.0004976	-5.68	0.000	-.0038036	-.0018524
bmi	.0006939	.0005693	1.22	0.223	-.0004222	.0018101
_cons	5.524951	.040588	136.12	0.000	5.445381	5.604522

- a. (5 points) Provide an interpretation for the intercept in the above regression model. What scientific use would you make of this estimate?

Ans: The exponentiated intercept $e^{5.525} = 250.9$ mg/dl is the estimated geometric mean cholesterol in newborns having no weight. There are no such people, and thus this number is not scientifically relevant. (You could have described the parameter as the estimated log geometric mean.)

- b. (5 points) Provide an interpretation for the slope for the age predictor in the above regression model. What scientific use would you make of this estimate?

Ans: The exponentiated slope parameter $e^{-0.002828} = 0.9972$ is the estimated ratio of geometric mean cholesterol between two populations differing in their ages by 1 year but agreeing in their BMI (with the older group tending toward the lower cholesterol level). This estimates the association between age and cholesterol after adjusting for BMI. (You could have described the parameter as the estimated log ratio of geometric means.)

- c. (5 points) Provide an interpretation for the slope for the bmi predictor in the above regression model. What scientific use would you make of this estimate?

Ans: The exponentiated slope parameter $e^{-0.0006939} = 1.0006941$ is the estimated ratio of geometric mean cholesterol between two populations differing in their BMI by 1 kg/m² but agreeing in their age (with the group having the higher BMI tending toward a higher cholesterol level. This estimates the association between BMI and cholesterol after adjusting for age. (You could have described the parameter as the estimated log ratio of geometric means.)

3. Now suppose that we are interested in the association between survival and cholesterol in this population. In addition to the variables mentioned in problem 1, we have measurements *ttodth*, which measures time in days to the earlier of death or follow-up, and *death*, an indicator that a death was observed. Consider the following logistic regression analysis.

. logistic death cholest bmi age

```
Logistic regression                Number of obs   =      4940
                                   LR chi2(3)         =      457.64
                                   Prob > chi2         =      0.0000
Log likelihood = -2399.4326        Pseudo R2       =      0.0871
```

death	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
cholest	.9951214	.0009306	-5.23	0.000	.993299	.996947
bmi	.9976596	.0078028	-0.30	0.764	.9824829	1.013071
age	1.129945	.0071683	19.26	0.000	1.115983	1.144082

- a. (10 points) What scientific conclusions do you reach from the above analysis regarding the association between survival and serum cholesterol level?

Ans: None. The variable *death* was measured over varying time periods. A proper analysis should take into account the censored observations as recorded in the variables *death* and *ttodth*. (Of note in this data set—though you had no way of knowing it—the subjects with the earliest censoring tend to be minorities. This is because a second cohort of subjects was recruited to enhance minority representation in the study. Failure to adjust for the censoring might therefore lead to very different estimates of survival probabilities to the extent that cholesterol levels, age, and BMI and/or their association with survival differed according to race/ethnicity. There are, of course, many nuances here, but the point remains: With censored observations you really need to consider that aspect.)

4. The following analyses present the overall survival probability for the sample at 2 years (730 days) and 4 years (1460 days), as well as a proportional hazards regression analysis.

```
. stset ttodth death
. sts list, at(730, 1461)
```

```
failure _d: death
analysis time _t: ttodth
```

Time	Beg. Total	Fail	Survivor Function	Std. Error	[95% Conf. Int.]	
730	4795	206	0.9588	0.0028	0.9529	0.9640
1461	4506	289	0.9010	0.0042	0.8924	0.9090

```
. stcox cholest, robust
```

```
failure _d: death
analysis time _t: ttodth
```

No. of subjects	=	4953	Number of obs	=	4953
No. of failures	=	1111			
Time at risk	=	11757827			
Log pseudolikelihood	=	-9173.1039	Wald chi2(1)	=	41.53
			Prob > chi2	=	0.0000

_t	Haz. Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
cholest	.9945688	.0008405	-6.44	0.000	.9929228	.9962175

- a. (10 points) Based on the above analyses, how would you answer the scientific question about associations between cholesterol and death from any cause? Provide a brief sentence suitable for a scientific report.

Ans: From a proportional hazards regression we estimate that when comparing two populations differing in their serum cholesterol levels there is a 0.543% decreased instantaneous risk of death per 1 mg/dl difference in serum cholesterol, with lower risk of death in populations having lower cholesterol (95% CI 0.708% lower to 0.378% lower). These results are beyond that that might be attributed to a chance observation in the absence of a true association ($P < 0.0005$).

- b. (10 points) Based on the above analysis, how would you characterize the clinical importance of any association between cholesterol and mortality?

Ans: The estimated 0.543% decreased instantaneous risk of death per 1 mg/dl difference in serum cholesterol translates to a clinically significant difference when we consider the variability of serum cholesterol within the sample. The interquartile range of 236 – 186 = 50 mg/dl would suggest that subjects at the 75th percentile of cholesterol levels would have an instantaneous risk of death of $0.99457^{50} = 0.7617$ times as high as that for subjects at the 25th percentile of cholesterol levels.

5. The following proportional hazards analysis also adjusted for age and bmi.

. stcox cholest age bmi, robust

```

failure _d: death
analysis time _t: ttodth

No. of subjects      =          4940          Number of obs      =          4940
No. of failures      =           1107
Time at risk         =       11728284

Log pseudolikelihood =      -8923.2449          Wald chi2(3)       =       545.38
                                                Prob > chi2        =       0.0000
    
```

_t	Haz. Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]
cholest	.995937	.0008132	-4.99	0.000	.9943445 .997532
age	1.107007	.0051113	22.02	0.000	1.097034 1.11707
bmi	1.003448	.0070465	0.49	0.624	.9897314 1.017354

- a. (5 points) How does the interpretation for the cholesterol slope parameter in this analysis differ from that in problem 2?

Ans: This analysis estimates a 0.406% decreased instantaneous risk of death per 1 mg/dl difference in serum cholesterol for two populations having similar age and BMI.

- b. (10 points) Does age confound the association between cholesterol and survival as considered in problem 4? Justify your answer.

Ans: Yes. Age is associated with survival independent of cholesterol level, with an estimated 10.7% increase in risk of death per 1 year difference in age between two groups. From the data presented in problem 1, there is an association between cholesterol and age in this data: Two groups differing in age by 20 years would be expected to differ in mean cholesterol levels by 10.9 mg/dl—a difference that might cause a spurious association between cholesterol and survival when one considers the impact of age on survival. (We also note that the adjusted HR is closer to the null than is the unadjusted HR. This would tell us that age and BMI were jointly confounding the cholesterol-survival association. Had the adjusted HR been further from the null, we could not have said anything about whether there was or was not confounding in this manner.)

- c. (10 points) Does BMI confound the association between cholesterol and survival as considered in problem 4? Justify your answer.

Ans: No. BMI does not appear to be associated with survival independent of cholesterol level in any substantial manner: The estimated 0.34% increase in risk of death per 1 kg/m² difference in BMI would only predict a maximal increase of 1.0034^{5.7} = 1.0195-fold over a difference equal to that between the 25th and 75th percentile of the BMI distribution. Furthermore, from the data presented in problem 1, there is not a substantial association between cholesterol and BMI in this data: Even two groups differing in BMI by 40 kg/m² (the range of BMI in this sample) would be expected to differ in mean cholesterol levels by 6.4 mg/dl—a difference that would not tend to cause a spurious association between cholesterol and survival given the lack of a substantial association between BMI and survival. (You received full credit if you just remarked on the lack of association between BMI and survival, or the lack of association between BMI and cholesterol. Either one of these conditions is sufficient to preclude confounding. However, if you justified your answer with a P value, you lost credit. P values play no role in assessing the association between the POI and the third variable in the sample, and a sufficiently high association between the POI and the third variable in the sample can cause sufficient variance inflation to preclude statistically significant effects in the adjusted analysis.)

6. The following analysis also added a predictor *cholsqr* = *cholest* ^ 2.

```
. stcox cholest cholsqr age bmi, robust
```

```

      failure _d: death
      analysis time _t: ttodth

No. of subjects      =          4940      Number of obs      =          4940
No. of failures      =           1107
Time at risk         =       11728284

Log pseudolikelihood =      -8920.7128      Wald chi2(4)        =          568.23
                                                Prob > chi2         =           0.0000
-----+-----
      |              Robust
      |              Std. Err.      z    P>|z|      [95% Conf. Interval]
-----+-----
cholest |      .9852513      .0045487   -3.22   0.001      .9763762      .9942071
cholsqr |      1.000025      .0000108    2.36   0.018      1.000004      1.000047
age      |      1.106517      .0051164   21.89   0.000      1.096534      1.11659
bmi      |      1.003524      .0070509    0.50   0.617      .9897994      1.017439
    
```

- a. (10 points) Based on the above analysis, describe how you would test for an association between serum cholesterol and probability of survival.

Ans: Because cholesterol is modeled using two covariates, we would have to test for an association between survival and cholesterol by a “multiple – partial” test simultaneously testing the slope parameters for *cholest* and *cholsqr*.

- b. (10 points) Based on the above analysis, is the effect of cholesterol on the log hazard rate well approximated by a straight line?

Ans: No. The P value testing for a zero slope of the quadratic term *cholsqr* is statistically significant ($P= 0.018$), so we can reject a hypothesis of a straight line relationship between the log hazard and cholesterol level.

- c. (Bonus: 15 points) Based on the above analysis, what can you say about the qualitative effect of cholesterol on the probability of survival? That is, how does elevated cholesterol affect the probability of survival?

Ans: We can compare the HR estimates comparing the reference population (age 0, BMI 0, cholesterol 0) to populations having a cholesterol equal the minimum (73 mg/dl), the median (210 mg /dl), and the maximum (430 mg/dl) for the sample (holding age and BMI each constant at 0—unrealistic values, but as we only care about the trends related to cholesterol, this is truly immaterial: We would obtain the same results holding age and BMI constant at any specific values). We thus compute a HR for a given value c of cholesterol as $.9852513^c \times 1.000025^{c \times c}$. For a cholesterol value of 73, we obtain $HR = 0.9852513^{73} \times 1.000025^{5329} = 0.3862$. For a cholesterol value of 210, we obtain $HR = 0.9852513^{210} \times 1.000025^{44100} = 0.1329$. For a cholesterol value of 430, we obtain $HR = 0.9852513^{430} \times 1.000025^{184900} = 0.1709$. Hence we see the prediction of a clear U-shaped trend, with lower risks of death for intermediate values of cholesterol than for extremely low or extremely high values. (Based on the above regression estimates, we find that the minimal risk of death is estimated to occur when the cholesterol value is approximately 297 mg/dl. While I would hesitate to really trust this exact value—I doubt the true “dose-response” is exactly a parabola—I will note that more flexible methods of modeling the cholesterol effect all tend to agree with the general result of a U-shaped trend having a minimum in the upper 200s.)

Distribution of grades:

Maximum possible (including bonus): 165

Mean	115.6
SD	17.8
10 %ile	91
20 %ile	99
30 %ile	103
40 %ile	109
50 %ile	117
60 %ile	122
70 %ile	127
80 %ile	131
90 %ile	136
Max	158